A MILITARILY OPERATIONAL AUTOMATIC INTERPRETING SYSTEM

*Melvyn Hunt**, Paul Bamberg*, Jay Tucker* & Steven Anderson**

**Dragon Systems UK
Research & Development Ltd
Millbank, Bishops Cleeve,
Cheltenham
Glos, England, GL52 4RW

*Dragon Systems, Inc
320 Nevada Street
Newton, MA 02160
USA

ABSTRACT

This paper describes a real-time interpreting system in which the operator speaks one of around 4000 phrases in one language, which is automatically recognised and the corresponding spoken phrase in the target language is played through a loudspeaker. This system has been used operationally by NATO forces. The basic system is first described, followed by an account of the wide range of uses to which this relatively simple one-way interpreting system can be put. Some developments of the basic system are then listed, both developments that are already in place and some that have potential for future implementation. Finally, an account is given of some relevant research on the use of statistical phonetic mapping techniques for extending the usability of such systems to non-native speakers of the source language.

1. BACKGROUND

This paper is concerned with a cross-language automatic interpreting system that has been used operationally by Nato forces. Experience with this system demonstrates, we feel, that relatively simple technology can perform a surprisingly useful task.

Reportedly, during the Gulf War Alliance forces had some difficulty in finding enough Arabic speakers to communicate with the very large numbers of Iraqi prisoners that had to be handled. This led the US military to seek some automated means of communicating with people in languages other than English. Dr. Lee Morin of the U.S. Navy created the "Medical Translator," with a point-and-click interface, to permit simple medical interviews. Anticipating a similar situation in Bosnia, DARPA asked Dragon System to add a voice interface to the Medical Translator, and Dragon Systems responded by developing the Multilingual Interview System, which first saw operational use in Serbo-Croatian with US forces assigned to the UN in Bosnia. More recently, in response to the troubles in Kosovo, an Albanian version has been developed.

2. THE MULTILINGUAL INTERVIEW SYSTEM

The first version of this system, which saw service in Bosnia, was based on the discrete-utterance, large-vocabulary speech recognition system, *DragonDictate*® [1, 2]. The term "discrete-utterance recogniser" is normally taken to be a synonym of "isolated-word recogniser". However, DragonDictate is capable of accepting and recognising long phrases. The first system developed had a "vocabulary" of 4000 such fixed phrases. They could be developed simply by providing the orthographic text of for each phrase and using the built-in 200,000-word pronouncing dictionary to develop a phonetic spelling for the phrase. Each phrase, no matter how long, was modeled as if it were an "isolated word."

The corresponding phrases in the target language are recorded by a native speaker of that language and stored as digitised waveforms. This provides a spoken output that is much more intelligible and natural than is possible with the current state-of-the-art in automatic text-to-speech systems. In any case, text-to-speech systems, or at least good-quality text-to-speech systems, exist only for a small number of major languages, not necessarily including the languages of interest for the Multilingual Interpreting System.

An operator speaks one of the 4000 phrases. He or she then confirms that the recogniser has correctly identified the phrase, either by seeing it displayed on a screen or — if eyes-free operation is needed — by having a recorded version of that phrase spoken back to the user. After confirmation, the phrase is then converted to the target language by simple table look-up, and the corresponding recorded phrase in the target language is played out through a loudspeaker. In cases where there exist several phrases that differ only in their final words, the system saves disk space by playing back a concatenation of two or more recordings, *e.g.* "I am a member" + "of the NATO peacekeeping forces."

Because the discrete-utterance recogniser needs to perform less computation than a continuous speech

recogniser, the hardware requirements are more modest. Portable computers using 486-style processors can be used in place of the Pentium-style processors needed for large-vocabulary continuous speech recognition, reducing the weight and our requirements of the portable equipment. The only technical weakness of the system is that it is vulnerable to errors in the "rapid match" portion of the discrete-utterance recogniser, which narrows the list of candidate phrases to 1000 or fewer by inspecting only the first 300 milliseconds of speech.

One of the operational systems was based on the *Fujitsu* portable PC, which thanks to speech recognition was particularly compact in that it needed no keyboard or mouse during use, its only input being via a headset-mounted microphone. This PC has the unusual feature of a monochrome transflective display that is easily readable in bright sunlight.

Although the system could be used with all 4000 phrases simultaneously active, it has often been found to be convenient to use the phrases in situation-specific subsets, such as those appropriate for a medical examination or for landmine clearance. Even the most dedicated user could not memorize all of the 4000 phrases, but individual users quickly learned the subset needed for their own tasks. The system included several techniques (categories, keyword search, prebuilt dialogues) to help users find phrases that were unfamiliar to them.

When used as a conventional dictation system, *DragonDictate* normally needs to be adapted to the voice of the user. However, in the Multilingual Interpreting System, especially when used with phrase subsets, it has generally been found to perform satisfactorily in speaker-independent mode, allowing military personnel of the same gender to share the same system freely without any need to signal to the system that a change of user has occurred.

Although this system was originally developed for operators whose language is American English, it could in principle be operated in several other major languages, since the *DragonDictate* recogniser on which it is based is available in British English, French, Italian, German, Spanish and Swedish. In practice, because of the length of the phrases and the vocabulary restriction, speakers of British English and indeed other national variants of English can satisfactorily operate a system set up for American English.

Generating a system for new target language is a simple operation, requiring the 4000 phrases to be translated into the new language and a native speaker of that language to record them.

3. USES OF THE MULTILINGUAL INTERVIEW SYSTEM

The system described in the previous section clearly operates only in one direction: from English into Serbo-Croatian, for example. This might appear at first to be a crippling limitation, since spoken communication is normally a two-way process. There are, of course, situations, such as crowd control, where one-way communication is all that is needed. But in a surprisingly large proportion of cases where two-way communication is needed, the Multilingual Interpreting System can perform a useful task. This section will describe just a few of the ways and situations in which it can be effective.

Often, questions can be posed in a way that allow yes/no responses. The military user can learn the words for "yes" and "no" in the target language or head movement gestures for these words may be common between the two languages. It is of course important to avoid negative questions (*e.g.* "You aren't injured, are you?) where the meaning of responses using the words normally translated as "yes" and "no" differs between languages.

In medical examinations, many things that need to be said are either instructions (*e.g.* "Please lie still", "Please open your mouth") or items of information (*e.g.* "I'm going to give you an injection to ease the pain"). The appropriate response to some others (*e.g.* "Point to where it hurts") is a gesture rather than a verbal response.

In gathering personal information, the individual being addressed can respond by writing down an answer (*e.g.* his or her date of birth, name, place of birth...). This will always be comprehensible for numerical information and for other information provided the language uses the Roman alphabet. Even when it does not use the Roman alphabet, the written information can be saved as a bitmap and taken away to be interpreted by others not necessarily located in the field of operation.

In the important area of avoiding or clearing minefields, individuals providing information can indicate locations on a map displayed on the computer screen and point to the kind of mine that has been laid when shown a screen that displays pictures of various mines. Such responses can be acted upon immediately or saved as annoted graphics for later review.

At security checks, the phrases being interpreted will normally be instructions, such as a request to leave a vehicle, to present identity papers or to hand over any firearms.

Finally, in cases where what is required is an extended verbal response, but the information required is not urgent, the person being interrogated can have his or her spoken response recorded for translation away from the

field of operation. The Multilingual Interview System is provided with the ability to make such recordings.

One of the advantages of the Multilingual Interview System surprisingly did not involve communication in the conventional sense at all. Reportedly, the use by soldiers of the system was a source of fascination to Bosnian young men, who were drawn into better relations with the soldiers because of it.

4. DEVELOPMENTS

The second version of the Multilingual Interview System allowed the recognition process to be enhanced from the fixed-phrase recogniser used in the first version to a true continuous large-vocabulary recogniser, namely the recogniser used in Dragon's general-purpose continuous speech recognition product, *Dragon NaturallySpeaking*™ [3]. At the price of requiring a more powerful microprocessor, this allows much greater flexibility in the form of the input in the source language. For example, a user does not have to remember whether "Please point to where it hurts" or "Point to where it hurts, please" is the required form of the phrase: both can be accepted with the more flexible arrangement that the continuous recogniser permits.

Both versions of the Multilingual Interview System have thus been based on a recogniser designed primarily for the very large vocabulary speech recognition needed for general-purpose dictation, where the grammar used must be probabilistic and allow in principle any word to occur in any context. Dragon Systems have recently developed a more compact recogniser suited to tasks in which the vocabulary and the structures of the phrases constructed from the vocabulary are more constrained. This will in principle allow the Multilingual Interview System to function on simpler hardware with lower power consumption yet with the flexibility provided by the second version just described.

Future developments can be envisaged that take advantage of research carried out at Dragon Systems on robust speech recognition in noisy conditions [4]. Currently, the operator of the Multilingual Interpreting System wears a headset-mounted close-talking microphone. Such a microphone contains a pressure-gradient element, making it relatively insensitive to distant sources of noise and consequently able to function well in high-noise environments. In some situations, however, it may be more natural and convenient to use a more conventional desk- or lapel-mounted microphone. Such microphones do not have the noise-cancelling properties of the headset-mounted microphone, but the developments in noise-robust recognition should allow the Multilingual Interpreting System to function with them in noisy conditions, at least when the noise is reasonably steady, such as noise from machinery.

The work on speech recognition in noise has also led to techniques for very rapid adaptation to the voices of native speakers of the source language and to techniques for compensating for the Lombard effect [5, 6] (the changes that occur in the voice of a user when the noise environment changes — principally an increase in the loudness of the speech when the noise gets louder).

5. POSSIBILITY OF USE WITH NON-NATIVE SPEAKERS

The performance of a system with non-native speakers is clearly of central interest to this workshop. Although no research using the Multilingual Interview System has been carried out with non-native speakers, relevant tests have been carried out in the framework of the development of speech recognition in noise just described [4].

The performance of the noise-robust recognition system was tested in noisy conditions in speaker-independent mode with both native and non-native speakers in a phrase recognition task not dissimilar from that in which the Multilingual Interview System might be used with a vocabulary of a few hundred words. The non-native speakers showed error rates roughly six times greater than the native speakers. The phonetic models used in recognition were then adapted using data from prompted training utterances from the speakers. There is no attempt to train the recogniser for specific words; rather, we use statistical techniques to map [7] the speaker's phonetic system into that of the standard language. In these particular tests, the training utterances were not chosen to give a balanced phonetic coverage of the task vocabulary but rather were selected randomly from phrases that can occur during use.

Figure 1 shows that adaptation is very effective with the non-native speakers, with a useful reduction in error rate after just 10 utterances, and a factor-of-three reduction after 80 utterances. The proportional reduction with native speakers is much less: only about 40% for female speakers and no clear improvement at all for male speakers. We have found in our tests that adaptation with our non-native test speakers always improved recognition performance, while with native speakers performance could even be degraded if an inadequate amount of adaptation material was used. This experience encourages our belief that the statistical techniques are indeed mapping acoustic realisations of phonemes from some consistent but non-standard forms produced by non-native speakers to something closer to standard forms. Note, however, that despite the evident effectiveness of the adaptation, the error rate with the non-native speakers remains about twice as high as the unadapted error rate with native speakers.

Of course, the term "non-native speaker" covers an immense range of deviation from the standard form of

the language, both in extent and in the type of deviation. Nevertheless, the adaptation behaviour seen here might reasonably be expected to be seen with any non-native speakers whose deviations from the norm correspond substantially to non-standard but consistent acoustic realisations of particular phonemes.

A key aspect of this kind of adaptation is that it is "supervised"; that is, the system knows what the speaker actually said. In the experiments just described this was achieved by prompting the speakers to produce the training utterances. In the Multilingual Interpreting System it is usual for the user to confirm that the system has correctly recognised the phrase spoken before it is translated into the target language. This process achieves the end of confirming to the system what the speaker actually said, and consequently adaptation of the form just described could be carried out unobtrusively during use.

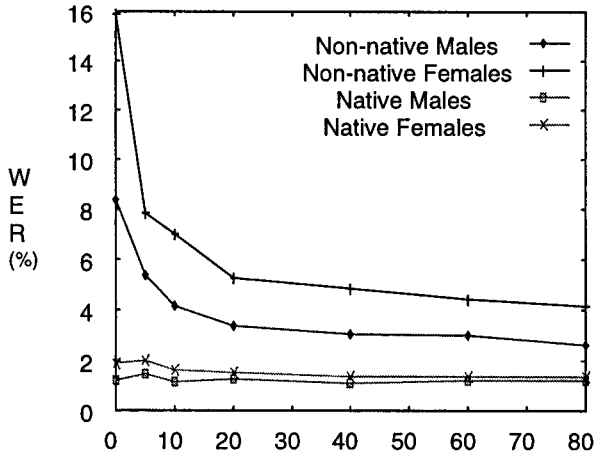

Figure 1 Word recognition error rate versus number of utterances used to adapt to the speakers for native and non-native speakers

6. CONCLUSIONS

This paper has attempted to show that an operationally useful automatic interpreting system for both military and non-military applications can be constructed from the current widely available large vocabulary speech recognition technology. The one-way nature of the interpreter does not prevent it from being effective in a wide range of tasks. There is much scope for the development of such a system to allow use with non-native speakers and in noisy environments without close-talking microphones, and for further reductions in the size and weight of the hardware required.

REFERENCES

1. J. Barnett, P. Bamberg, M. Held, J. Huerta, L. Manganaro, A. Weiss, "Comparative Recognition Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages", *Eurospeech '95*, vol. I, pp. 189-192.

2. J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, B. Peskin, "Multilingual Speech Recognition at Dragon Systems", *ICSLP '96*, pp. 2191-2194.

3. R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann, J. Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer", *Proc. ARPA Spoken Language Systems Tech. Workshop*, Austin, 1995, pp. 116-119.

4. M. J. Hunt, "Some Experience in In-Car Speech Recognition" *Proc. IEEE/Nokia Workshop on Robust Methods for Speech Recognition in Adverse Conditions.* May 25-26, 1999, Tampere, Finland, pp. 25-32.

5. E. Lombard "Le Signe de l'Elevation de la Voix", *Ann. Maladies Orielle, Larynx, Nez, Pharynx*, Vol 37, 1911, pp. 101-119

6. J-C Junqua, "The influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon Known as the Lombard Reflex", *Speech Communication*, 1996, Vol. 20, pp. 13-22.

7. V. Nagesha & L. Gillick, "Studies in Transformation-Based Adaptation", *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-97,* Munich, Germany, April 1997, Vol. II, pp. 1031-1034.